FACULTY OF DATA AND DECISION SCIENCES
MAKING SENSE OF IT

הטכניון
מכון טכנולוגי לישראל
TECHNION
Israel Institute of Technology

Evaluating Reasoning in Vision-Language Models (VLM)

Shachaf Haviv, Shani Angel, Saar Manshrov

Motivation

Vision-Language Models (VLMs) struggle to perform nuanced reasoning tasks that require understanding relationships between images and context.

This project aims to advance Visual-Temporal Reasoning in VLMs by adapting the abductive Natural Language Inference (NLI) task to the visual domain. The goal is to push the boundaries of VLMs' reasoning capabilities using a meticulously crafted dataset designed to rigorously test and challenge the models' ability to infer plausible scenarios from visual information. To further explore the capabilities of VLMs, we selected three distinct image styles to assess whether the models exhibit varying levels of inference performance across different visual contexts.

Premise

"A pregnant woman stands in the kitchen"

Hypothesis 1

"A woman stands in the kitchen, holding a baby."

Hypothesis 2

"A woman stands in the kitchen, holding a watermelon."

Data Creation

Textual Data Source
The textual dataset consists of 64 examples, with 56 carefully selected from an existing dataset to ensure diverse representation across reasoning categories, including logical, physical, and functional. Additionally, 8 new examples were handcrafted to further enrich the dataset's variety and depth.

Image Generation and Photography
A total of **600** images were generated using MidJourney for the 64 examples, each represented in three distinct styles: **Hand drawn sketch**, **Japanese Anime**, and **Paper Cut**. For 17 of the examples, real photographs were taken, adding an extra Photographed style to further enhance the dataset.

Data Set

Anime | Paper Cut | Sketched | Photorealistic | Photographed

Premise
A couple sits on a couch with a couples therapist sitting opposite them

Hypothesis 1
A couple in the kitchen arguing and shouting at each other

Hypothesis 2
A couple in the kitchen hugging each other

Category VS Domain Distribution

	cultural	emotional	functional	logical	physical	social	Total
business & work		1		1		2	**4**
daily life & hospitality	7	3	5	6	6	5	**32**
education		1		1	2		**4**
healthcare		1	3	4		1	**9**
religion	3						**3**
science & technology					1	1	**2**
sports & arts		3	2			1	**6**
transportation		1		2	1		**4**
Total	**10**	**10**	**10**	**14**	**10**	**10**	**64**

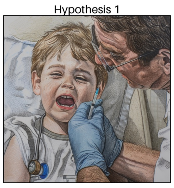
In addition to our custom dataset, we integrated an external datasets contributed by another team and Mor Ventura, which introduced four additional styles. This expansion enriched the diversity of visual representations used for model evaluation, bringing the total number of samples to 685.

Model Evaluation

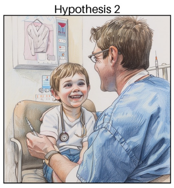
Model Reasoning Demo

Premise

Hypothesis 1



Hypothesis 2



Openbmb MiniCPM Llama3 Answer

*"The first image is of a child who appears to be crying or in discomfort, with an adult administering what seems to be a medical procedure. The second image shows a smiling child who seems to be at ease and engaged in a playful activity. Based on the context provided, **the first image is more plausible** as it aligns with the description of a child receiving a vaccine or similar medical treatment, which often results in the child crying due to the pain and discomfort associated with such procedures."*

Openbmb MiniCPM Llama3 Reversed Answer

*"The first image is of a child with a slight smile and eyes looking away from the camera, while receiving an injection from an adult. The second image shows a child crying and pulling away during an injection. Based on the context provided, **the first image is more plausible** as it depicts a child who seems to be handling the situation with less distress than the child in the second image."*

Models and Prompts

HuggingFaceM4_idefics2-8b

This is the context image: <image>
This is scenario 1: <image>
This is scenario 2: <image>
based on the context
image, which scenario is
more plausible

google_paligemma-3b-mix-224

You are presented with 3 images. Given the leftmost image (context), which one of the other two optional scenes is more plausible? Answer only 'middle' or 'right.'

openbmb_MiniCPM-Llama3-V-2_5-int4

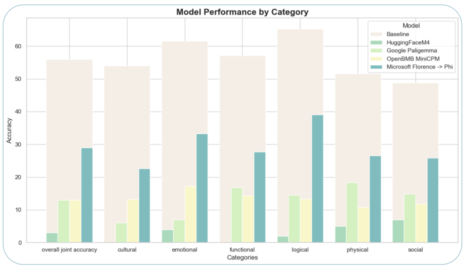
This is the context image: <image> This is the first plausible image: <image> This is the second plausible image: <image> Describe in a sentence each image. Based on the context, which one of the other two images is more plausible? Answer "first" or "second."

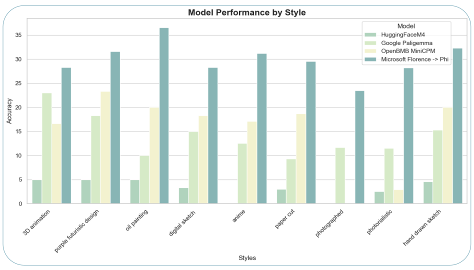
microsoft-Florence-2-large (image to text) -> microsoft-Phi-3-mini-4k (text to text)

image to text prompt:
<DETAILED_CAPTION>
text to text prompt: First scenario: ..
Second scenario: ...
Given the context: which scenario is more plausible?

Joint Accuracy

The models exhibited bias toward the order of image presentation. To address this, we tested each model with the images presented in reverse order. The joint accuracy shown, combines the results from both the original and reversed image sequences.





Conclusions

Our evaluation of Vision-Language Models (VLMs) showed that most models struggled to accurately assess the plausibility of hypotheses based on images, often favoring the first image presented.

Having the model describe the images before making a judgment appears to enhance its decision-making accuracy.

Photorealistic and photographed image styles do not seem to provide any advantages over other styles; in fact, they may perform worse.

Limited computational resources constrained our ability to test models with very large numbers of parameters, potentially affecting the results.